不对称约束多人非零和博弈的自适应评判控制

李梦花, 王　鼎, 乔俊飞†

(北京工业大学 信息学部, 北京 100124; 计算智能与智能系统北京市重点实验室, 北京 100124;

智慧环保北京实验室, 北京 100124; 北京人工智能研究院, 北京 100124)

摘要: 本文针对连续时间非线性系统的不对称约束多人非零和博弈问题, 建立了一种基于神经网络的自适应评判控制方法. 首先, 本文提出了一种新颖的非二次型函数来处理不对称约束问题, 并且推导出最优控制律和耦合Hamilton-Jacobi方程. 值得注意的是, 当系统状态为零时, 最优控制策略是不为零的, 这与以往不同. 然后, 通过构建单一评判网络来近似每个玩家的最优代价函数, 从而获得相关的近似最优控制策略. 同时, 在评判学习期间发展了一种新的权值更新规则. 此外, 通过利用Lyapunov理论证明了评判网络权值近似误差和闭环系统状态的稳定性. 最后, 仿真结果验证了本文所提方法的有效性.

关键词: 神经网络; 自适应评判控制; 自适应动态规划; 非线性系统; 不对称约束; 多人非零和博弈

引用格式: 李梦花, 王鼎, 乔俊飞. 不对称约束多人非零和博弈的自适应评判控制. 控制理论与应用, 2023, 40(9): 1562 – 1568

DOI: 10.7641/CTA.2022.20063

Adaptive critic control for multi-player non-zero-sum games with asymmetric constraints

LI Meng-hua, WANG Ding, QIAO Jun-fei†

(Faculty of Information Technology, Beijing University of Technology, Beijing 100124, China;
Beijing Key Laboratory of Computational Intelligence and Intelligent System, Beijing 100124, China;
Beijing Laboratory of Smart Environmental Protection, Beijing 100124, China;
Beijing Institute of Artificial Intelligence, Beijing 100124, China)

Abstract: In this paper, an adaptive critic control method based on the neural networks is established for multi-player non-zero-sum games with asymmetric constraints of continuous-time nonlinear systems. First, a novel nonquadratic function is proposed to deal with asymmetric constraints, and then the optimal control laws and the coupled Hamilton-Jacobi equations are derived. It is worth noting that the optimal control strategies do not stay at zero when the system state is zero, which is different from the past. After that, only a critic network is constructed to approximate the optimal cost function for each player, so as to obtain the associated approximate optimal control strategies. Meanwhile, a new weight updating rule is developed during critic learning. In addition, the stability of the weight estimation errors of critic networks and the closed-loop system state is proved by utilizing the Lyapunov method. Finally, simulation results verify the effectiveness of the method proposed in this paper.

Key words: neural networks; adaptive critic control; adaptive dynamic programming; nonlinear systems; asymmetric constraints; multi-player non-zero-sum games

Citation: LI Menghua, WANG Ding, QIAO Junfei. Adaptive critic control for multi-player non-zero-sum games with asymmetric constraints. *Control Theory & Applications*, 2023, 40(9): 1562 – 1568

1　引言

自适应动态规划(adaptive dynamic programming, ADP)方法由Werbos[1]首先提出, 该方法结合了动态规划、神经网络和强化学习, 其核心思想是利用函数近似结构来估计最优代价函数, 从而获得被控系统的近似最优解. 在ADP方法体系中, 动态规划蕴含最优

收稿日期: 2022−01−21; 录用日期: 2022−11−10.

†通信作者. E-mail: adqiao@bjut.edu.cn.

本文责任编委: 王龙.

科技创新2030 –“新一代人工智能”重大项目(2021ZD0112302, 2021ZD0112301), 国家重点研发计划项目(2018YFC1900800–5), 北京市自然科学基金项目(JQ19013), 国家自然科学基金项目(62222301, 61890930–5, 62021003)资助.

Supported by the National Key Research and Development Program of China (2021ZD0112302, 2021ZD0112301, 2018YFC1900800–5), the Beijing Natural Science Foundation (JQ19013) and the National Natural Science Foundation of China (62222301, 61890930–5, 62021003).

性原理提供理论基础, 神经网络作为函数近似结构提供实现手段, 强化学习提供学习机制. 值得注意的是, ADP方法具有强大的自学习能力, 在处理非线性复杂系统的最优控制问题上具有很大的潜力[2–7]. 此外, ADP作为一种近似求解最优控制问题的新方法, 已经成为智能控制与计算智能领域的研究热点. 关于ADP的详细理论研究以及相关应用, 读者可以参考文献[8–9]. 本文将基于ADP的动态系统优化控制统称为自适应评判控制.

近年来, 微分博弈问题在控制领域受到了越来越多的关注. 微分博弈为研究多玩家系统的协作、竞争与控制提供了一个标准的数学框架, 包括二人零和博弈、多人零和博弈以及多人非零和博弈等. 在零和博弈问题中, 控制输入试图最小化代价函数而干扰输入试图最大化代价函数. 在非零和博弈问题中, 每个玩家都独立地选择一个最优控制策略来最小化自己的代价函数. 值得注意的是, 零和博弈问题已经被广泛研究. 在文献[10]中, 作者提出了一种改进的ADP方法来求解多输入非线性连续系统的二人零和博弈问题. An等人[11]提出了两种基于积分强化学习的算法来求解连续时间系统的多人零和博弈问题. Ren等人[12]提出了一种新颖的同步脱策方法来处理多人零和博弈问题. 然而, 关于非零和博弈[13–14]的研究还很少. 此外, 控制约束在实际应用中也广泛存在. 这些约束通常是由执行器的固有物理特性引起的, 如气压、电压和温度. 因此, 为了确保被控系统的性能, 受约束的系统需要被考虑. Zhang等人[15]发展了一种新颖的事件采样ADP方法来求解非线性连续约束系统的鲁棒最优控制问题. Huo等人[16]研究了一类非线性约束互联系统的分散事件触发控制问题. Yang和He[17]研究了一类具有不匹配扰动和输入约束的非线性系统事件触发鲁棒镇定问题. 这些文献考虑的都是对称约束, 而实际应用中, 被控系统受到的约束也可能是不对称的[18–20], 例如在污水处理过程中, 需要通过氧传递系数和内回流量对溶解氧浓度和硝态氮浓度进行控制, 而根据实际的运行条件, 这两个控制变量就需要被限制在一个不对称约束范围内[20]. 因此, 在控制器设计过程中, 不对称约束问题将是笔者研究的一个方向.

到目前为止, 关于具有控制约束的微分博弈问题, 有一些学者取得了相应的研究成果[12,21–23]. 但可以发现, 具有不对称约束的多人非零和博弈问题还没有学者研究. 同时, 在多人非零和博弈问题中, 相关的耦合Hamilton-Jacobi(HJ)方程是很难求解的. 因此, 本文针对一类连续时间非线性系统的不对称约束多人非零和博弈问题, 提出了一种自适应评判控制方法来近似求解耦合HJ方程, 从而获得被控系统的近似最优解. 本文的主要贡献如下: 1) 首次将不对称约束应用到连续时间非线性系统的多人非零和博弈问题中; 2) 提出了一种新颖的非二次型函数来处理不对称约束问题, 并且当系统状态为零时, 最优控制策略是不为零的, 这与以往不同; 3) 在学习期间, 用单一评判网络结构代替了传统的执行–评判网络结构, 并且提出了一种新的权值更新规则; 4) 利用Lyapunov方法证明了评判网络权值近似误差和系统状态的一致最终有界 (uniformly ultimately bounded, UUB) 稳定性.

2　问题描述

考虑以下具有不对称约束的N–玩家连续时间非线性系统:

$$\dot{x}(t)=f(x(t))+\sum_{j=1}^{N}g_j(x(t))u_j(t), \tag{1}$$

其中: $x(t)\in\Omega\subset\mathbb{R}^n$是状态向量且$x(0)=x_0$为初始状态, $\mathbb{R}^n$代表由所有n-维实向量组成的欧氏空间, Ω是$\mathbb{R}^n$的一个紧集; $u_j(t)\in\mathfrak{T}_j\subset\mathbb{R}^m$为玩家$j$在时刻$t$所选择的策略, 且$\mathfrak{T}_j$为

$$\begin{aligned}\mathfrak{T}_j=\{&[u_{j1}\ \ u_{j2}\ \ \cdots\ \ u_{jm}]^{\mathrm{T}}\in\mathbb{R}^m: u^j_{\min}\leqslant u_{jl}\leqslant u^j_{\max},\\ &|u^j_{\min}|\neq|u^j_{\max}|, l=1,2,\cdots,m\},\end{aligned} \tag{2}$$

其中: $u^j_{\min}\in\mathbb{R}$和$u^j_{\max}\in\mathbb{R}$分别代表控制输入分量的最小界和最大界, $\mathbb{R}$表示所有实数集.

假设 1　非线性系统(1)是可控的, 并且$x=0$是被控系统 (1) 的一个平衡点. 此外, $\forall j\in\mathbb{N}$, $f(x)$和$g_j(x)$是未知的Lipschitz函数且$f(0)=0$, 其中集合$\mathbb{N}=\{1,2,\cdots,N\}$, $N\geqslant 2$是一个正整数.

假设 2　$\forall j\in\mathbb{N}$, $g_j(0)=0$, 且存在一个正常数b_{g_j}使$\|g_j(x)\|\leqslant b_{g_j}$, 其中$\|\cdot\|$表示在$\mathbb{R}^n$上的向量范数或者在$\mathbb{R}^{n\times m}$上的矩阵范数, $\mathbb{R}^{n\times m}$ 代表由所有$n\times m$维实矩阵组成的空间.

注 1　假设1–3是自适应评判领域的常用假设, 例如文献[6, 13, 19], 是为了保证系统的稳定性以及方便后文中的稳定性证明, 其中假设3出现在后文中的第3.2节.

定义与每个玩家相关的效用函数为

$$U_i(x,\mathfrak{U})=x^{\mathrm{T}}Q_ix+\sum_{j=1}^{N}\mathcal{S}_j(u_j),\ i\in\mathbb{N}, \tag{3}$$

其中$\mathfrak{U}=\{u_1,u_2,\cdots,u_N\}$并且$Q_i$是一个对称正定矩阵. 此外, 为了处理不对称约束问题, 令$\mathcal{S}_j(u_j)$为

$$\mathcal{S}_j(u_j)=2\alpha_j\sum_{l=1}^{m}\int_{\beta_j}^{u_{jl}}\tanh^{-1}(\frac{\mathfrak{z}-\beta_j}{\alpha_j})\mathrm{d}\mathfrak{z}, \tag{4}$$

其中α_j和β_j分别为

$$\alpha_j=\frac{u^j_{\max}-u^j_{\min}}{2},\ \beta_j=\frac{u^j_{\max}+u^j_{\min}}{2}. \tag{5}$$

因此, 与每个玩家相关的代价函数可以表示为

$$J_i(x_0,\mathfrak{U})=\int_0^{\infty}U_i(x,\mathfrak{U})\mathrm{d}\tau,\ i\in\mathbb{N}, \tag{6}$$

本文希望构建一个Nash均衡$\mathfrak{U}^*=\{u_1^*,u_2^*,\cdots,u_N^*\}$, 来使以下不等式被满足:

$$J_i(u_1^*,\cdots,u_i^*,\cdots,u_N^*)\leqslant J_i(u_1^*,\cdots,u_i,\cdots,u_N^*),\tag{7}$$

其中$i\in\mathbb{N}$. 为了方便, 将$J_i(x_0,\mathfrak{U})$简写为$J_i(x_0)$. 于是, 每个玩家的最优代价函数为

$$J_i^*(x_0)=\min_{u_i}J_i(x_0,\mathfrak{U}),\ i\in\mathbb{N}.\tag{8}$$

在本文中, 如果一个控制策略集的所有元素都是可容许的, 那么这个集合是可容许的.

定义 1 (容许控制[24]) 如果控制策略$u_i(x)$是连续的, $u_i(x)$可以镇定系统(1), 并且$J_i(x_0)$是有限的, 那么它是集合Ω上关于代价函数(6)的可容许控制律, 即$u_i(x)\in\Psi(\Omega),i\in\mathbb{N}$, 其中, $\Psi(\Omega)$是Ω上所有容许控制律的集合.

对于任意一个可容许控制律$u_i(x)\in\Psi(\Omega)$, 如果相关代价函数(6)是连续可微的, 那么非线性Lyapunov方程为

$$0=U_i(x,\mathfrak{U})+(\nabla J_i(x))^{\mathrm{T}}(f(x)+\sum_{j=1}^{N}g_j(x)u_j),\tag{9}$$

其中: $i\in\mathbb{N}$, $J_i(0)=0$, 并且$\nabla(\cdot)\triangleq\dfrac{\partial(\cdot)}{\partial x}$. 根据最优控制理论, 耦合HJ方程为

$$0=\min_{\mathfrak{U}}H_i(x,\mathfrak{U},\nabla J_i^*(x)),i\in\mathbb{N},\tag{10}$$

其中, Hamiltonian函数$H_i(x,\mathfrak{U},\nabla J_i^*(x))$为

$$H_i(x,\mathfrak{U},\nabla J_i^*(x))=U_i(x,\mathfrak{U})+(\nabla J_i^*(x))^{\mathrm{T}}(f(x)+\sum_{j=1}^{N}g_j(x)u_j),\tag{11}$$

进而, 由$\dfrac{\partial H_i(x,\mathfrak{U},\nabla J_i^*(x))}{\partial u_i}=0$可得出最优控制律为

$$u_i^*(x)=-\alpha_i\tanh(\frac{1}{2\alpha_i}g_i^{\mathrm{T}}(x)\nabla J_i^*(x))+\bar{\beta}_i,\ i\in\mathbb{N},\tag{12}$$

其中$\bar{\beta}_i=[\beta_i\ \ \beta_i\ \ \cdots\ \ \beta_i]^{\mathrm{T}}\in\mathbb{R}^m$.

注 2 根据式(2)和式(5), 能推导出$\beta_i\neq0$, 即$\bar{\beta}_i\neq0$, 又根据式(12)可知$u_i^*(0)\neq0,i\in\mathbb{N}$. 因此, 为了保证$x=0$是系统(1)的平衡点, 在假设2中提出了条件$\forall j\in\mathbb{N},g_j(0)=0$.

将式(12)代入式(10), 耦合HJ方程又能表示为

$$(\nabla J_i^*(x))^{\mathrm{T}}f(x)+\sum_{j=1}^{N}((\nabla J_i^*(x))^{\mathrm{T}}g_j(x)\bar{\beta}_j)+x^{\mathrm{T}}Q_ix-\sum_{j=1}^{N}((\nabla J_i^*(x))^{\mathrm{T}}\alpha_jg_j(x)\tanh(A_j(x)))+\sum_{j=1}^{N}\mathcal{S}_j(-\alpha_j\tanh(A_j(x))+\bar{\beta}_j)=0,\ i\in\mathbb{N},\tag{13}$$

其中$J_i^*(0)=0$并且$A_j(x)=\dfrac{1}{2\alpha_j}g_j^{\mathrm{T}}(x)\nabla J_j^*(x)$.

如果已知每个玩家的最优代价函数值, 那么相关的最优状态反馈控制律就可以直接获得, 也就是说式(13)是可解的. 可是, 式(13)这种非线性偏微分方程的求解是十分困难的. 同时, 随着系统维数的增加, 存储量和计算量也随之以指数形式增加, 也就是平常所说的"维数灾"问题. 因此, 为了克服这些弱点, 在第3部分提出了一种基于神经网络的自适应评判机制, 来近似每个玩家的最优代价函数, 从而获得相关的近似最优状态反馈控制策略.

3 自适应评判控制设计

3.1 神经网络实现

本节的核心是构建并训练评判神经网络, 以得到训练后的权值, 从而获得每个玩家的近似最优代价函数值.

首先, 根据神经网络的逼近性质[25], 可将每个玩家的最优代价函数$J_i^*(x)$在紧集Ω上表示为

$$J_i^*(x)=W_i^{\mathrm{T}}\sigma_i(x)+\xi_i(x),\ i\in\mathbb{N},\tag{14}$$

其中: $W_i\in\mathbb{R}^\delta$是理想权值向量, $\sigma_i(x)\in\mathbb{R}^\delta$是激活函数, δ是隐含层神经元个数, $\xi_i(x)\in\mathbb{R}$是重构误差. 同时, 可得出每个玩家的最优代价函数梯度为

$$\nabla J_i^*(x)=(\nabla\sigma_i(x))^{\mathrm{T}}W_i+\nabla\xi_i(x),\ i\in\mathbb{N},\tag{15}$$

将式(15)代入式(12), 有

$$u_i^*(x)=-\alpha_i\tanh(B_i(x)+C_i(x))+\bar{\beta}_i,\ i\in\mathbb{N},\tag{16}$$

其中: $B_i(x)=\dfrac{1}{2\alpha_i}g_i^{\mathrm{T}}(x)(\nabla\sigma_i(x))^{\mathrm{T}}W_i\in\mathbb{R}^m$, $C_i(x)=\dfrac{1}{2\alpha_i}g_i^{\mathrm{T}}(x)\nabla\xi_i(x)\in\mathbb{R}^m$. 然后, 将式(15)代入式(13), 耦合HJ方程变为

$$\begin{aligned}&W_i^{\mathrm{T}}\nabla\sigma_i(x)f(x)+(\nabla\xi_i(x))^{\mathrm{T}}f(x)+x^{\mathrm{T}}Q_ix+\\&\sum_{j=1}^{N}((W_i^{\mathrm{T}}\nabla\sigma_i(x)+(\nabla\xi_i(x))^{\mathrm{T}})g_j(x)\bar{\beta}_j)-\\&\sum_{j=1}^{N}(\alpha_jW_i^{\mathrm{T}}\nabla\sigma_i(x)g_j(x)\tanh(B_j(x)+C_j(x)))-\\&\sum_{j=1}^{N}(\alpha_j(\nabla\xi_i(x))^{\mathrm{T}}g_j(x)\tanh(B_j(x)+C_j(x)))+\\&\sum_{j=1}^{N}\mathcal{S}_j(-\alpha_j\tanh(B_j(x)+C_j(x))+\bar{\beta}_j)=0,\ i\in\mathbb{N}.\end{aligned}\tag{17}$$

值得注意的是, 式(14)中的理想权值向量W_i是未知的, 也就是说式(16)中的$u_i^*(x)$是不可解的. 因此,

构建如下的评判神经网络:

$$\hat{J}_i^*(x)=\hat{W}_i^{\mathrm{T}}\sigma_i(x),\ i\in\mathbb{N}, \tag{18}$$

来近似每个玩家的最优代价函数, 其中$\hat{W}_i\in\mathbb{R}^\delta$是估计的权值向量. 同时, 其梯度为

$$\nabla\hat{J}_i^*(x)=(\nabla\sigma_i(x))^{\mathrm{T}}\hat{W}_i,\ i\in\mathbb{N}. \tag{19}$$

考虑式(19), 近似的最优控制律为

$$\hat{u}_i^*(x)=-\alpha_i\tanh(D_i(x))+\bar{\beta}_i,\ i\in\mathbb{N}, \tag{20}$$

其中 $D_i(x)=\dfrac{1}{2\alpha_i}g_i^{\mathrm{T}}(x)(\nabla\sigma_i(x))^{\mathrm{T}}\hat{W}_i$. 同理, 近似的Hamiltonian可以写为

$$\begin{aligned}&\hat{H}_i(x,\hat{W}_i)=\\&\hat{W}_i^{\mathrm{T}}\phi_i+x^{\mathrm{T}}Q_ix+\sum_{j=1}^N(\hat{W}_i^{\mathrm{T}}\nabla\sigma_i(x)g_j(x)\bar{\beta}_j)-\\&\sum_{j=1}^N(\alpha_j\hat{W}_i^{\mathrm{T}}\nabla\sigma_i(x)g_j(x)\tanh(D_j(x)))+\\&\sum_{j=1}^N\mathcal{S}_j(-\alpha_j\tanh(D_j(x))+\bar{\beta}_j),\ i\in\mathbb{N},\end{aligned} \tag{21}$$

其中$\phi_i=\nabla\sigma_i(x)f(x)$. 此外, 定义误差量 $e_i=\hat{H}_i(x,\hat{W}_i)-H_i(x,\mathfrak{U}^*,\nabla J_i^*(x))=\hat{H}_i(x,\hat{W}_i)$. 为了使$e_i$足够小, 需要训练评判网络来使目标函数$E_i=\dfrac{1}{2}e_i^{\mathrm{T}}e_i$最小化. 在这里, 本文采用的训练准则为

$$\begin{aligned}\dot{\hat{W}}_i=&-\gamma_i\frac{1}{(1+\phi_i^{\mathrm{T}}\phi_i)^2}\left(\frac{\partial E_i}{\partial\hat{W}_i}\right)=\\&-\gamma_i\frac{\phi_i}{(1+\phi_i^{\mathrm{T}}\phi_i)^2}e_i,\ i\in\mathbb{N},\end{aligned} \tag{22}$$

其中: $\gamma_i>0$是评判网络的学习率, $(1+\phi_i^{\mathrm{T}}\phi_i)^2$用于归一化操作. 此外, 定义评判网络的权值近似误差为$\tilde{W}_i=W_i-\hat{W}_i$. 因此, 有

$$\dot{\tilde{W}}_i=\gamma_i\frac{\varphi_i}{1+\phi_i^{\mathrm{T}}\phi_i}e_{Hi}-\gamma_i\varphi_i\varphi_i^{\mathrm{T}}\tilde{W}_i,\ i\in\mathbb{N}, \tag{23}$$

其中: $\varphi_i=\dfrac{\phi_i}{(1+\phi_i^{\mathrm{T}}\phi_i)}$, $e_{Hi}=-(\nabla\xi_i(x))^{\mathrm{T}}f(x)$是残差项.

3.2　稳定性分析

本节的核心是通过利用Lyapunov方法讨论评判网络权值近似误差和闭环系统状态的UUB稳定性. 这里, 给出以下假设:

假　设 3　$\|\nabla\xi_i(x)\|\leqslant b_{\nabla\xi_i}$, $\|\nabla\sigma_i(x)\|\leqslant b_{\nabla\sigma_i}$, $\|e_{Hi}\|\leqslant b_{e_{Hi}}$, $\|W_i\|\leqslant b_{W_i}$, 其中: $b_{\nabla\xi_i}$, $b_{\nabla\sigma_i}$, $b_{e_{Hi}}$, b_{W_i}都是正常数, $i\in\mathbb{N}$.

定理 1　考虑系统(1), 如果假设1–3成立, 状态反馈控制律由式(20)给出, 且评判网络权值通过式(22)进行训练, 则评判网络权值近似误差$\tilde{W}_i$是UUB稳定的.

证　选取如下的Lyapunov函数:

$$L_1(t)=\sum_{i=1}^N(\frac{1}{2}\tilde{W}_i^{\mathrm{T}}\tilde{W}_i)=\sum_{i=1}^N L_{1i}(t), \tag{24}$$

计算$L_{1i}(t)$沿着式(23)的时间导数, 即

$$\begin{aligned}\dot{L}_{1i}(t)=&\gamma_i\tilde{W}_i^{\mathrm{T}}\frac{\varphi_i}{1+\phi_i^{\mathrm{T}}\phi_i}e_{Hi}-\\&\gamma_i\tilde{W}_i^{\mathrm{T}}\varphi_i\varphi_i^{\mathrm{T}}\tilde{W}_i,\ i\in\mathbb{N},\end{aligned} \tag{25}$$

利用不等式 $\bar{X}^{\mathrm{T}}\bar{Y}\leqslant\dfrac{1}{2}\|\bar{X}\|^2+\dfrac{1}{2}\|\bar{Y}\|^2$ (注: $\bar{X}$和$\bar{Y}$都是具有合适维数的向量), 并且考虑$1+\phi_i^{\mathrm{T}}\phi_i\geqslant 1$, 能得到

$$\begin{aligned}\dot{L}_{1i}(t)\leqslant&\frac{\gamma_i}{2}(\|\varphi_i^{\mathrm{T}}\tilde{W}_i\|^2+\|e_{Hi}\|^2)-\gamma_i\tilde{W}_i^{\mathrm{T}}\varphi_i\varphi_i^{\mathrm{T}}\tilde{W}_i=\\&-\frac{\gamma_i}{2}\tilde{W}_i^{\mathrm{T}}\varphi_i\varphi_i^{\mathrm{T}}\tilde{W}_i+\frac{\gamma_i}{2}\|e_{Hi}\|^2,\ i\in\mathbb{N}.\end{aligned} \tag{26}$$

根据假设3, 有

$$\dot{L}_{1i}(t)\leqslant-\frac{\gamma_i}{2}\lambda_{\min}(\varphi_i\varphi_i^{\mathrm{T}})\|\tilde{W}_i\|^2+\frac{\gamma_i}{2}b_{e_{Hi}}^2,\ i\in\mathbb{N}, \tag{27}$$

其中$\lambda_{\min}(\cdot)$表示矩阵的最小特征值. 因此, 当不等式

$$\|\tilde{W}_i\|>\sqrt{\frac{b_{e_{Hi}}^2}{\lambda_{\min}(\varphi_i\varphi_i^{\mathrm{T}})}},\ i\in\mathbb{N} \tag{28}$$

成立时, 有$\dot{L}_{1i}(t)<0$. 根据标准的Lyapunov定理[26], 可知评判网络权值近似误差$\tilde{W}_i$是UUB稳定的.

证毕.

定理 2　考虑系统(1), 如果假设1–3成立, 状态反馈控制律由式(20)给出, 且评判网络权值通过式(22)进行训练, 则系统状态$x(t)$是UUB稳定的.

证　选取如下的Lyapunov函数:

$$L_{2i}(t)=J_i^*(x),\ i\in\mathbb{N}. \tag{29}$$

计算$L_{2i}(t)$沿着系统$\dot{x}=f(x)+\sum\limits_{j=1}^N g_j(x)\hat{u}_j^*$的时间导数, 即

$$\begin{aligned}\dot{L}_{2i}(t)=&(\nabla J_i^*(x))^{\mathrm{T}}(f(x)+\sum_{j=1}^N g_j(x)\hat{u}_j^*)=\\&(\nabla J_i^*(x))^{\mathrm{T}}(f(x)+\sum_{j=1}^N g_j(x)u_j^*)+\\&\sum_{j=1}^N((\nabla J_i^*(x))^{\mathrm{T}}g_j(x)(\hat{u}_j^*-u_j^*)),\ i\in\mathbb{N}.\end{aligned} \tag{30}$$

考虑式(13), 有

$$\begin{aligned}\dot{L}_{2i}(t)=&-x^{\mathrm{T}}Q_ix-\sum_{j=1}^N\mathcal{S}_j(u_j^*)+\\&\underbrace{\sum_{j=1}^N((\nabla J_i^*(x))^{\mathrm{T}}g_j(x)(\hat{u}_j^*-u_j^*))}_{\Sigma_i},\ i\in\mathbb{N},\end{aligned} \tag{31}$$

利用不等式 $\bar{X}^{\mathrm{T}}\bar{Y} \leqslant \frac{1}{2}\|\bar{X}\|^2 + \frac{1}{2}\|\bar{Y}\|^2$, 并且考虑式(15)–(16)(20), 可得

$$\begin{aligned}&\Sigma_i \leqslant\\&\frac{1}{2}\sum_{j=1}^{N}\| -\alpha_j\tanh(D_j(x)) + \alpha_j\tanh(F_j(x))\|^2 +\\&\frac{1}{2}\sum_{j=1}^{N}\|g_j^{\mathrm{T}}(x)((\nabla\sigma_i(x))^{\mathrm{T}}W_i+\nabla\xi_i(x))\|^2,\ i\in\mathbb{N},\end{aligned} \tag{32}$$

其中$F_j(x)=B_j(x)+C_j(x)$. 然后, 利用不等式$\|\bar{X}+\bar{Y}\|^2 \leqslant 2\|\bar{X}\|^2 + 2\|\bar{Y}\|^2$, 有

$$\begin{aligned}\Sigma_i \leqslant& \sum_{j=1}^{N}(\|\alpha_j\tanh(D_j(x))\|^2+\|\alpha_j\tanh(F_j(x))\|^2)+\\&\sum_{j=1}^{N}\|g_j^{\mathrm{T}}(x)(\nabla\sigma_i(x))^{\mathrm{T}}W_i\|^2 +\\&\sum_{j=1}^{N}\|g_j^{\mathrm{T}}(x)\nabla\xi_i(x)\|^2,\ i\in\mathbb{N},\end{aligned} \tag{33}$$

其中 $D_j(x)\in\mathbb{R}^m$, $F_j(x)\in\mathbb{R}^m$ 分别被表示为 $[D_{j1}(x)\ D_{j2}(x)\ \cdots\ D_{jm}(x)]^{\mathrm{T}}$和$[F_{j1}(x)\ F_{j2}(x)\ \cdots\ F_{jm}(x)]^{\mathrm{T}}$. 易知, $\forall\theta\in\mathbb{R}, \tanh^2\theta \leqslant 1$. 因此, 有

$$\|\tanh(D_j(x))\|^2 = \sum_{l=1}^{m}\tanh^2(D_{jl}(x)) \leqslant m, \tag{34}$$

$$\|\tanh(F_j(x))\|^2 = \sum_{l=1}^{m}\tanh^2(F_{jl}(x)) \leqslant m. \tag{35}$$

同时, 根据假设2–3, 有

$$\Sigma_i \leqslant \sum_{j=1}^{N}(2\alpha_j^2 m + b_{g_j}^2 b_{\nabla\sigma_i}^2 b_{W_i}^2 + b_{g_j}^2 b_{\nabla\xi_i}^2),\ i\in\mathbb{N}, \tag{36}$$

根据式(2)(4)–(5), 可知$\mathcal{S}_j(u_j^*) \geqslant 0$. 于是, 有

$$\dot{L}_{2i}(t) \leqslant -\lambda_{\min}(Q_i)\|x\|^2 + \varpi_i,\ i\in\mathbb{N}, \tag{37}$$

其中 $\varpi_i = \sum_{j=1}^{N}(2\alpha_j^2 m + b_{g_j}^2 b_{\nabla\sigma_i}^2 b_{W_i}^2 + b_{g_j}^2 b_{\nabla\xi_i}^2)$. 因此, 根据式(37)可知, 当不等式$\|x\| > \sqrt{\dfrac{\varpi_i}{\lambda_{\min}(Q_i)}}$成立时, 有$\dot{L}_{2i}(t) < 0$. 即, 如果$x(t)$满足下列不等式:

$$\|x\| > \max\{\sqrt{\frac{\varpi_1}{\lambda_{\min}(Q_1)}}, \cdots, \sqrt{\frac{\varpi_N}{\lambda_{\min}(Q_N)}}\}, \tag{38}$$

则, $\forall i\in\mathbb{N}$, 都有$\dot{L}_{2i}(t) < 0$. 同理, 可得闭环系统状态$x(t)$也是UUB稳定的. 证毕.

4 仿真结果

考虑如下的3–玩家连续时间非线性系统:

$$\begin{aligned}\dot{x} =& \begin{bmatrix}-1.2x_1 + 1.5x_2\sin x_2\\ 0.5x_1 - x_2\end{bmatrix} +\\&\begin{bmatrix}0\\ 1.5\sin x_1\cos x_1\end{bmatrix}u_1(x) +\\&\begin{bmatrix}0\\ 1.2\sin x_1\cos x_2\end{bmatrix}u_2(x) +\\&\begin{bmatrix}0\\ 1.1\sin x_2\end{bmatrix}u_3(x),\end{aligned} \tag{39}$$

其中: $x(t)=[x_1\ \ x_2]^{\mathrm{T}}\in\mathbb{R}^2$是状态向量, $u_1(x)\in\mathfrak{T}_1=\{u_1\in\mathbb{R}: -1\leqslant u_1\leqslant 2\}$, $u_2(x)\in\mathfrak{T}_2=\{u_2\in\mathbb{R}: -0.2\leqslant u_2\leqslant 1\}$和$u_3(x)\in\mathfrak{T}_3=\{u_3\in\mathbb{R}: -0.4\leqslant u_3\leqslant 0.8\}$是控制输入.

令$Q_1=2I_2$, $Q_2=1.8I_2$, $Q_3=0.3I_2$, 其中I_2代表2×2维单位矩阵. 同时, 根据式(5)可知, $\alpha_1=1.5$, $\beta_1=0.5$, $\alpha_2=0.6$, $\beta_2=0.4$, $\alpha_3=0.6$, $\beta_3=0.2$. 因此, 与每个玩家相关的代价函数可以表示为

$$J_i(x_0) = \int_0^{\infty}(x^{\mathrm{T}}Q_i x + \sum_{j=1}^{3}\mathcal{S}_j(u_j))\mathrm{d}\tau,\ i=1,2,3, \tag{40}$$

其中

$$\begin{aligned}\mathcal{S}_j(u_j) =& 2\alpha_j\int_{\beta_j}^{u_j}\tanh^{-1}(\frac{\mathfrak{z}-\beta_j}{\alpha_j})\mathrm{d}\mathfrak{z} =\\&2\alpha_j(u_j-\beta_j)\tanh^{-1}(\frac{u_j-\beta_j}{\alpha_j}) +\\&\alpha_j^2\ln(1-\frac{(u_j-\beta_j)^2}{\alpha_j^2}).\end{aligned} \tag{41}$$

然后, 本文针对系统(39)构建3个评判神经网络, 每个玩家的评判神经网络权值分别为$\hat{W}_1=[\hat{W}_{11}\ \ \hat{W}_{12}\ \ \hat{W}_{13}]^{\mathrm{T}}$, $\hat{W}_2=[\hat{W}_{21}\ \ \hat{W}_{22}\ \ \hat{W}_{23}]^{\mathrm{T}}$, $\hat{W}_3=[\hat{W}_{31}\ \ \hat{W}_{32}\ \ \hat{W}_{33}]^{\mathrm{T}}$, 激活函数被定义为$\sigma_1(x)=\sigma_2(x)=\sigma_3(x)=[x_1^2\ \ x_1x_2\ \ x_2^2]^{\mathrm{T}}$, 且隐含层神经元个数为$\delta=3$. 此外, 系统初始状态取$x_0=[0.5\ \ -0.5]^{\mathrm{T}}$, 每个评判神经网络的学习率分别为$\gamma_1=1.5$, $\gamma_2=0.8$, $\gamma_3=0.2$, 且每个评判神经网络的初始权值都在0和2之间选取. 最后, 引入探测噪声$\eta(t)=\sin^2(-1.2t)\cos(0.5t)+\cos(2.4t)\sin^3(2.4t)+\sin^5 t+\sin^2(1.12t)+\sin^2 t\times\cos t+\sin^2(2t)\cos(0.1t)$, 使得系统满足持续激励条件.

执行学习过程, 本文发现每个玩家的评判神经网络权值分别收敛于$[6.9091\ \ 2.9904\ \ 6.6961]^{\mathrm{T}}$, $[4.8901\ \ 2.2347\ \ 5.2062]^{\mathrm{T}}$, $[1.7945\ \ 0.3321\ \ 2.4583]^{\mathrm{T}}$. 在60个时间步之后去掉探测噪声, 每个玩家的评判网络权值收敛过程如图1–3所示. 然后, 将训练好的权值代入式(20), 能得到每个玩家的近似最优控制律, 将其应用到系统(39), 经过10个时间步之后, 得到的状态轨迹和控制轨迹分别如图4–5所示. 由图4可知, 系统状态最终收敛到了平衡点. 由图5可知, 每个玩家的控制轨迹都没有超出预定的边界, 并且可以观察到u_1, u_2和u_3分别收敛于0.5, 0.4和0.2. 综上所述, 仿真结果验证了所提方法的有效性.

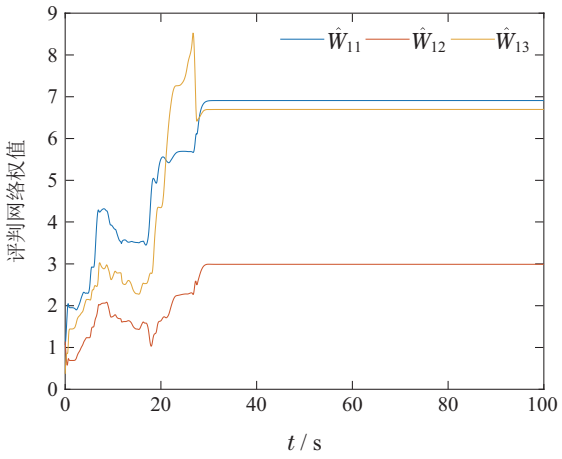

图 1　玩家1的评判网络权值收敛过程

Fig. 1　Convergence process of the critic network weights for player 1

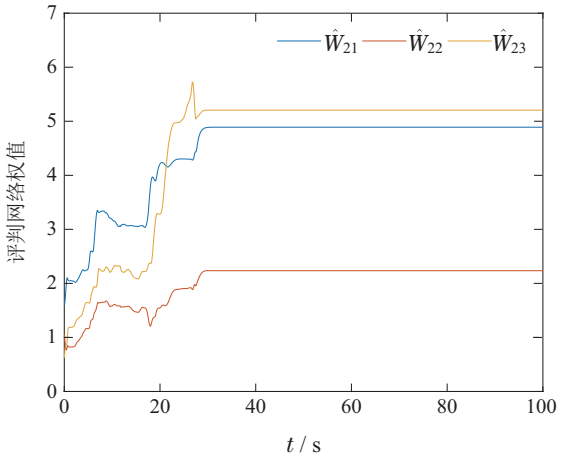

图 2　玩家2的评判网络权值收敛过程

Fig. 2　Convergence process of the critic network weights for player 2

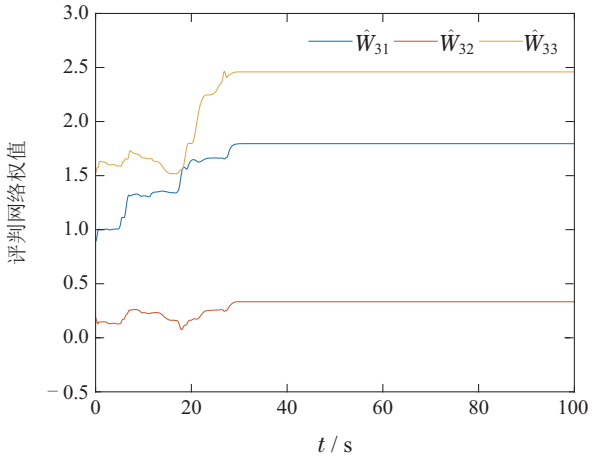

图 3　玩家3的评判网络权值收敛过程

Fig. 3　Convergence process of the critic network weights for player 3

5　结论

本文首次将不对称约束应用到连续时间非线性系统的多人非零和博弈问题中. 首先, 获得了最优状态反馈控制律和耦合HJ方程, 并且为了解决不对称约束问题, 建立了一种新的非二次型函数. 值得注意的是, 当系统状态为零时, 最优控制策略是不为零的. 其次, 由于耦合HJ方程不易求解, 提出了一种基于神经网络的自适应评判算法来近似每个玩家的最优代价函数, 从而获得相关的近似最优控制律. 在实现过程中, 用单一评判网络结构代替了经典的执行–评判结构, 并且建立了一种新的权值更新规则. 然后, 利用Lyapunov理论讨论了评判网络权值近似误差和系统状态的UUB稳定性. 最后, 仿真结果验证了所提算法的可行性. 在未来的工作中, 会考虑将事件驱动机制引入到连续时间非线性系统的不对称约束多人非零和博弈问题中, 并且将该研究内容应用到污水处理系统中也是笔者的一个重点研究方向.

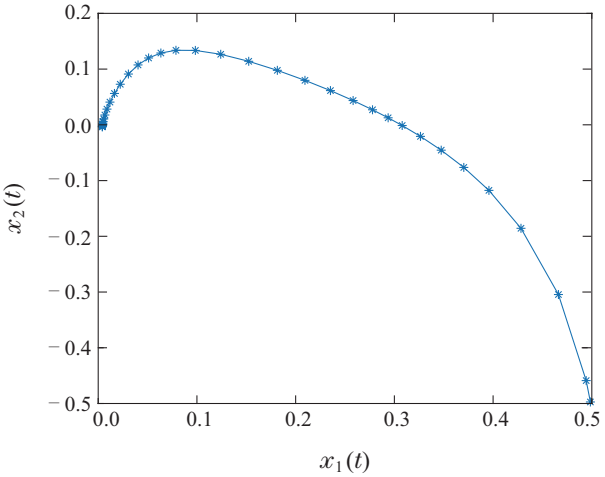

图 4　系统(39)的状态轨迹

Fig. 4　State trajectory of the system (39)

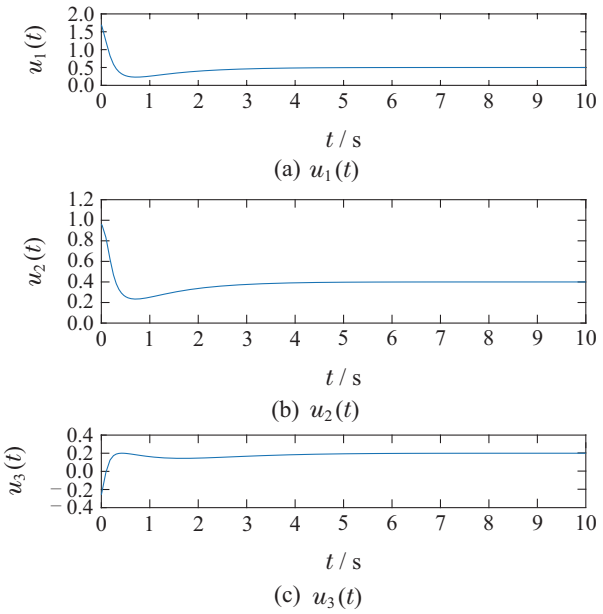

图 5　系统(39)的控制轨迹

Fig. 5　Control trajectories of the system (39)

参考文献:

[1] WERBOS P J. *Beyond regression: New tools for prediction and analysis in the behavioral sciences*. Cambridge: Harvard University, 1974.

[2] HONG Chengwen, FU Yue. Nonlinear robust approximate optimal tracking control based on adaptive dynamic programming. *Control Theory & Applications*, 2018, 35(9): 1285 – 1292.
(洪成文, 富月. 基于自适应动态规划的非线性鲁棒近似最优跟踪控制. 控制理论与应用, 2018, 35(9): 1285 – 1292.)

[3] CUI Lili, ZHANG Yong, ZHANG Xin. Event-triggered adaptive dynamic programming algorithm for the nonlinear zero-sum differential games. *Control Theory & Applications*, 2018, 35(5): 610 – 618.
(崔黎黎, 张勇, 张欣. 非线性零和微分对策的事件触发自适应动态规划算法. 控制理论与应用, 2018, 35(5): 610 – 618.)

[4] WANG D, HA M, ZHAO M. The intelligent critic framework for advanced optimal control. *Artificial Intelligence Review*, 2022, 55(1): 1 – 22.

[5] WANG D, QIAO J, CHENG L. An approximate neuro-optimal solution of discounted guaranteed cost control design. *IEEE Transactions on Cybernetics*, 2022, 52(1): 77 – 86.

[6] YANG X, HE H. Adaptive dynamic programming for decentralized stabilization of uncertain nonlinear large-scale systems with mismatched interconnections. *IEEE Transactions on Systems, Man, and Cybernetics: Systems*, 2020, 50(8): 2870 – 2882.

[7] ZHAO B, LIU D. Event-triggered decentralized tracking control of modular reconfigurable robots through adaptive dynamic programming. *IEEE Transactions on Industrial Electronics*, 2020, 67(4): 3054 – 3064.

[8] WANG Ding. Research progress on learning-based robust adaptive critic control. *Acta Automatica Sinica*, 2019, 45(6): 1037 – 1049.
(王鼎. 基于学习的鲁棒自适应评判控制研究进展. 自动化学报, 2019, 45(6): 1037 – 1049.)

[9] ZHANG Huaguang, ZHANG Xin, LUO Yanhong, et al. An overview of research on adaptive dynamic programming. *Acta Automatica Sinica*, 2013, 39(4): 303 – 311.
(张化光, 张欣, 罗艳红, 等. 自适应动态规划综述. 自动化学报, 2013, 39(4): 303 – 311.)

[10] LÜ Yongfeng, TIAN Jianyan, JIAN Long, et al. Approximate-dynamic-programming H_∞ controls for multi-input nonlinear system. *Control Theory & Applications*, 2021, 38(10): 1662 – 1670.
(吕永峰, 田建艳, 菅垄, 等. 非线性多输入系统的近似动态规划H_∞控制. 控制理论与应用, 2021, 38(10): 1662 – 1670.)

[11] AN P, LIU M, WAN Y, et al. Multi-player H_∞ differential game using on-policy and off-policy reinforcement learning. *The 16th International Conference on Control and Automation*. Electr Network: IEEE, 2020, 10: 1137 – 1142.

[12] REN H, ZHANG H, MU Y, et al. Off-policy synchronous iteration IRL method for multi-player zero-sum games with input constraints. *Neurocomputing*, 2020, 378: 413 – 421.

[13] LIU D, LI H, WANG D. Online synchronous approximate optimal learning algorithm for multiplayer nonzero-sum games with unknown dynamics. *IEEE Transactions on Systems, Man, and Cybernetics: Systems*, 2014, 44(8): 1015 – 1027.

[14] VAMVOUDAKIS K G, LEWIS F L. Non-zero sum games: Online learning solution of coupled Hamilton-Jacobi and coupled Riccati equations. *IEEE International Symposium on Intelligent Control*. Denver, CO, USA: IEEE, 2011, 9: 171 – 178.

[15] ZHANG H, ZHANG K, XIAO G, et al. Robust optimal control scheme for unknown constrained-input nonlinear systems via a plug-n-play event-sampled critic-only algorithm. *IEEE Transactions on Systems, Man, and Cybernetics: Systems*, 2020, 50(9): 3169 – 3180.

[16] HUO X, KARIMI H R, ZHAO X, et al. Adaptive-critic design for decentralized event-triggered control of constrained nonlinear interconnected systems within an identifier-critic framework. *IEEE Transactions on Cybernetics*, 2022, 52(8): 7478 – 7491.

[17] YANG X, HE H. Event-triggered robust stabilization of nonlinear input-constrained systems using single network adaptive critic designs. *IEEE Transactions on Systems, Man, and Cybernetics: Systems*, 2020, 50(9): 3145 – 3157.

[18] WANG L, CHEN C L P. Reduced-order observer-based dynamic event-triggered adaptive NN control for stochastic nonlinear systems subject to unknown input saturation. *IEEE Transactions on Neural Networks and Learning Systems*, 2021, 32(4): 1678 – 1690.

[19] YANG X, ZHU Y, DONG N, et al. Decentralized event-driven constrained control using adaptive critic designs. *IEEE Transactions on Neural Networks and Learning Systems*, 2022, 33(10): 5830 – 5844.

[20] WANG D, ZHAO M, QIAO J. Intelligent optimal tracking with asymmetric constraints of a nonlinear wastewater treatment system. *International Journal of Robust and Nonlinear Control*, 2021, 31(14): 6773 – 6787.

[21] LI M, WANG D, QIAO J, et al. Neural-network-based self-learning disturbance rejection design for continuous-time nonlinear constrained systems. *Proceedings of the 40th Chinese Control Conference*. Shanghai, China: IEEE, 2021, 7: 2179 – 2184.

[22] SU H, ZHANG H, JIANG H, et al. Decentralized event-triggered adaptive control of discrete-time nonzero-sum games over wireless sensor-actuator networks with input constraints. *IEEE Transactions on Neural Networks and Learning Systems*, 2020, 31(10): 4254 – 4266.

[23] YANG X, HE H. Event-driven H_∞-constrained control using adaptive critic learning. *IEEE Transactions on Cybernetics*, 2021, 51(10): 4860 – 4872.

[24] ABU-KHALAF M, LEWIS F L. Nearly optimal control laws for nonlinear systems with saturating actuators using a neural network HJB approach. *Automatica*, 2005, 41(5): 779 – 791.

[25] HORNIK K, STINCHCOMBE M, WHITE H. Universal approximation of an unknown mapping and its derivatives using multilayer feedforward networks. *Neural Networks*, 1990, 3(5): 551 – 560.

[26] LEWIS F L, JAGANNATHAN S, YESILDIREK A. *Neural Network Control of Robot Manipulators and Nonlinear Systems*. London: Taylor & Francis, 1999.

作者简介:

李梦花 博士研究生, 目前研究方向为自适应动态规划、智能控制, E-mail: limenghua@emails.bjut.edu.cn;

王 鼎 教授, 博士生导师, 目前研究方向为智能控制、强化学习, E-mail: dingwang@bjut.edu.cn;

乔俊飞 教授, 博士生导师, 目前研究方向为智能计算、智能优化控制, E-mail: adqiao@bjut.edu.cn.